1. Ranger是什么

Apache Ranger 是 Hadoop 平台上操作、监控、管理数据安全的集中式安全管理框架。Ranger 的愿景是在 Apache Hadoop 生态系统中提供全面的安全性。

目前，Apache Ranger 支持以下 Apache 项目的细粒度授权和审计：

```
Apache Hadoop/HDFS
Apache Hadoop/YARN
Apache Hive
Apache HBase
Apache Storm
Apache Knox
Apache Solr
Apache Kafka
Apache Nifi
```

等等其他组件。

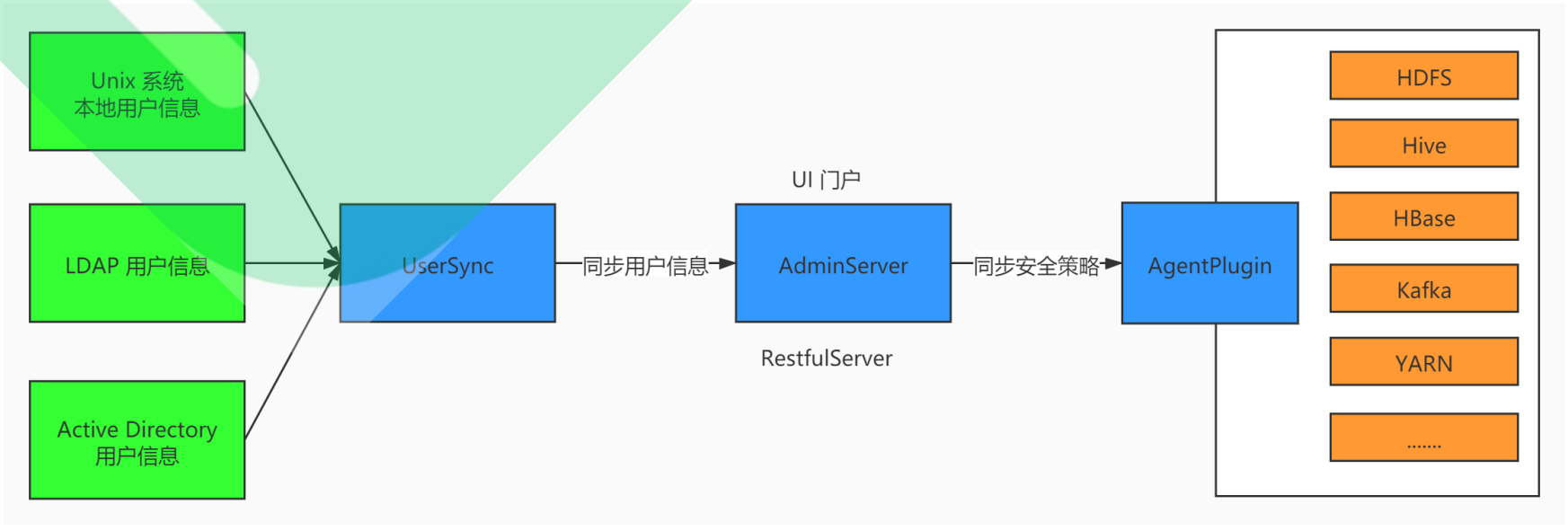
Apache Ranger 通过访问控制策略提供了一套标准的授权方法，改变了 Hadoop 上各个组件各自为政分散管理权限的现状。作为标准，Ranger 提供了集中式的组件，用于审计用户的访问行为和管理组件间的安全交互行为。

Apache Ranger 意为游侠，以下是它的官网和 logo：http://ranger.apache.org/

2. Ranger架构

2.1. 基本架构

Ranger 的基本架构如下图所示：



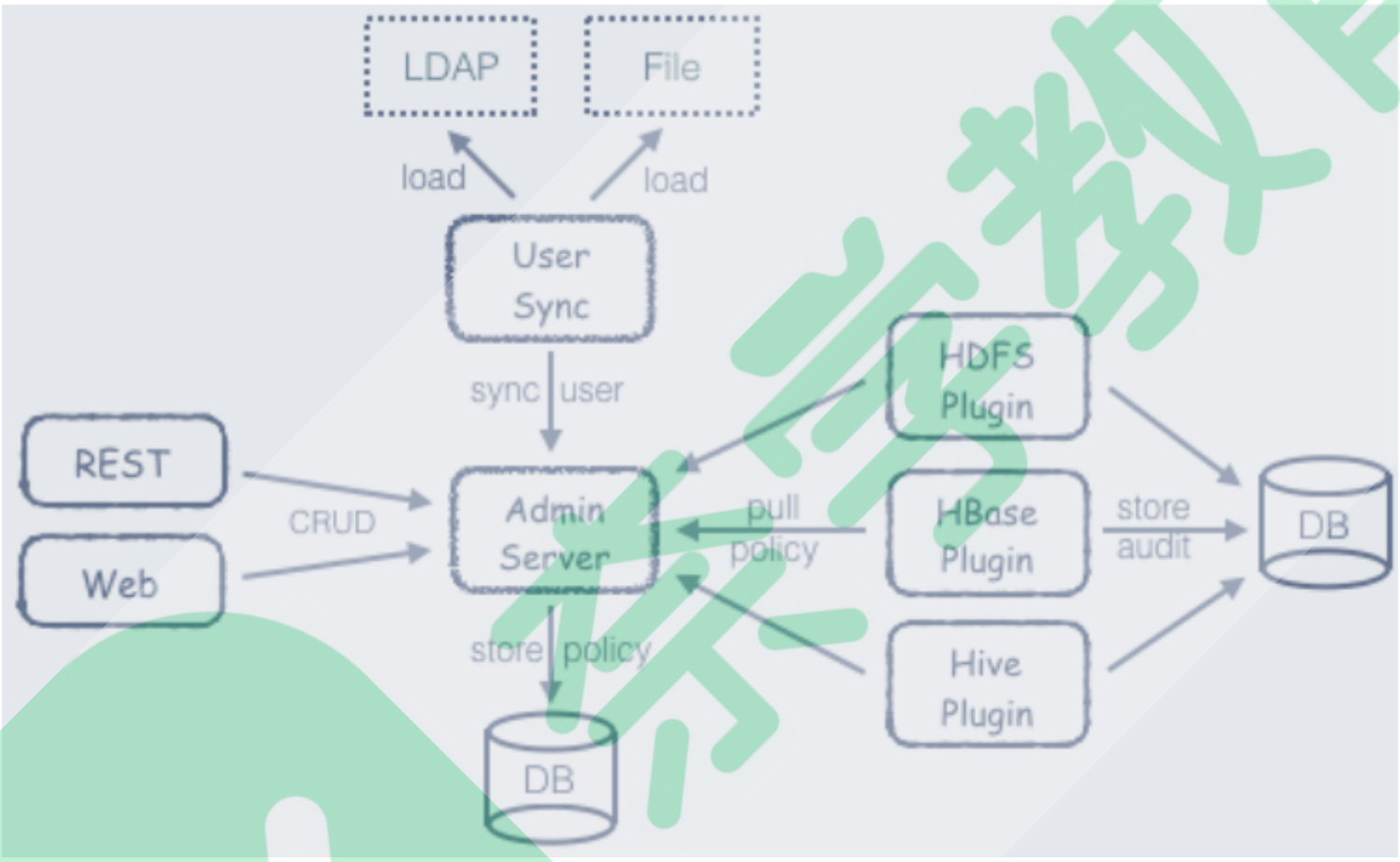
组件	作用
Ranger UserSync	定期将 Unix 系统或 LDAP 或 Active Directory 中的用户/组信息同步到 RangerAdmin 中。也可用作 RangerAdmin 的身份验证服务器，以使用 Linux 用户/密码登录到 RangerAdmin。
Ranger Admin Server	用于管理安全策略、用户/组的 UI 门户并提供 Rest Server。一般简称 RangerAdmin
Agent Plugin	插件是嵌入到 Hadoop 各个组件的轻量级 Java 程序。插件定期从 AdminServer 拉取策略，存储在本地文件中。当用户访问 Hadoop 组件时，插件会拦截请求根据策略进行安全评估，并且定期发送数据到审计服务器做记录。

关于两个名词的解释：

- LDAP：Light Directory Access Portocol，它是基于 X.500 标准的轻量级目录访问协议。
- Active Directory：Active Directory（活动目录）是微软 Windows Server 中，负责架构中大型网路环境的集中式目录管理服务（Directory Services），Windows 2000 Server 开始内建于 Windows Server 产品中，它处理了在组织中的网路物件，物件可以是**计算机，用户，群组，组织单元（OU）**等等，只要是在 Active Directory 结构定义档（schema）中定义的物件，就可以储存在 Active Directory 资料档中，并利用 Active Directory Service Interface 来存取。

2.2. 工作过程

工作过程如下：



2.3. Ranger 核心特性

Apache Ranger 的一些核心特性和设计目标：

```
1、集中安全管理，在中央 UI 或使用 REST API 管理所有与安全相关的任务。

2、精细授权，使用 Hadoop 组件/工具 执行特定动作或操作，并通过集中管理工具进行管理。

3、标准化所支持 Hadoop 组件的授权方法。

4、增强了对不同授权方法的支持：基于角色的访问控制，基于属性的访问控制，基于 Tag 的访问控制(需结合Atlas)等

5、在所支持的 Hadoop 组件中集中审计用户访问和管理操作（与安全相关）

6、支持和 Kerberos 的集成
```

3. 大数据安全方案

业界比较常见的大数据安全方案主要有三种：

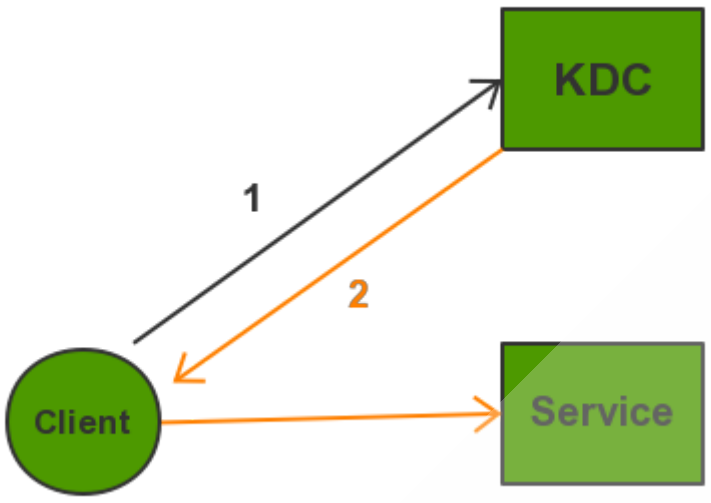
```
1、Kerberos（开源常用方案，业界比较常用的方案）

2、Apache Sentry（Cloudera 选用的方案，CDH 版本中集成，CDP 中已经换成了Ranger）

3、Apache Ranger（Hortonworks 选用的方案，HDP 发行版中集成）
```

3.1. kerberos

Kerberos 是一种基于对称密钥的身份认证协议，它作为一个独立的第三方的身份认证服务，可以为其它服务提供身份认证功能，且支持 SSO(即客户端身份认证后，可以访问多个服务如 HBase / HDFS 等)。



服务	作用
KDC	Kerberos 的服务端程序，用于验证各个模块
Client	需要访问服务的用户，KDC 和 Service 会对用户的身份进行认证
Service	即集成了 Kerberos 的服务，如 HDFS/YARN/HBase 等

Kerberos 协议过程主要有三个阶段：

1. 第一个阶段 Client 向 KDC 申请 TGT（Ticket Granting Ticket，认购权）
2. 第二个阶段 Client 通过获得的 TGT 向 KDC 申请用于访问 Service 的 Ticket
3. 第三个阶段 Client 用返回的 Ticket 访问 Service

优点：

```
1、服务认证，防止 broker datanode regionserver 等组件冒充加入集群
2、解决了服务端到服务端的认证，也解决了客户端到服务端的认证
```

缺点：

```
1、kerberos 为了安全性使用临时 ticket，认证信息会失效，用户多数情况下重新认证繁琐
2、kerberos 只能控制你访问或者拒绝访问一个服务，不能控制到很细的粒度，比如 HDFS 的某一个路径，Hive 的某一个表，对用户级别上的认证并没有实现(需要配合LDAP)
```

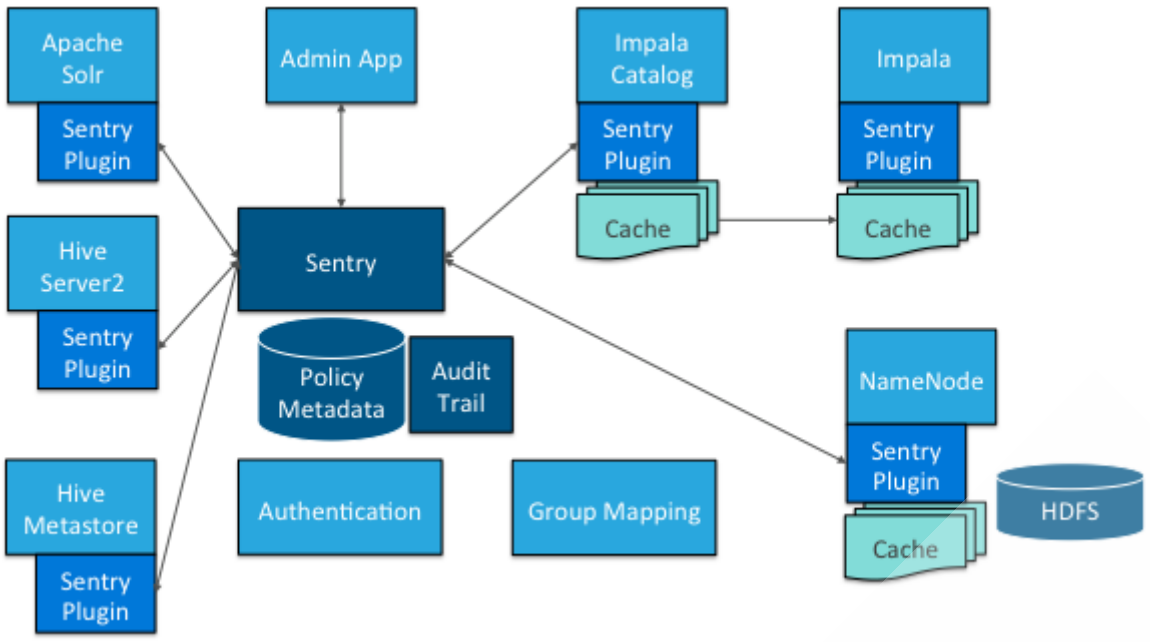
总结：

```
Kerberos 更专注的是通用的认证解决方案，但是它对授权以及其他安全方面的管理功能并不擅长、例如细粒度权限、审计等。往往跟 Ranger 或者 Sentry 结合使用各取所长。
```

3.2. Apache Sentry

Apache Sentry 是 Hadoop 之上的基于角色的细粒度授权模块。它为在 Hadoop 集群（特别是 CDH）上运行的经过身份验证的用户和应用程序提供数据访问授权。目前 Sentry 已支持 Apache Hive，Apache Solr，Apache Kafka，Apache Impala 和 HDFS (仅限于 Hive 表的 HDFS 数据权限同步)。

Sentry 是基于角色的，所以你在使用 Sentry 时你需要创建 Role，然后通过 Role 映射到 OS 或者 AD 中的 Group，然后再映射到访问 Hadoop 的最终用户。你可以使用 Sentry 来限制用户对 DB、TABLE、COLUMN 或 URI 的访问，权限设定可以通过 Hive 或 Impala 的命令行接口执行 Sentry 相关命令来实现，有关这些命 令的更多详细信息，请参见 Cloudera 的 Sentry 文档。



优点：

```
1、Sentry 支持细粒度的 HDFS 元数据访问控制，对 hive 支持列级别的访问控制
2、Sentry 通过基于角色的授权简化了管理，将访问同一数据集的不同特权级别授予多个角色
3、Sentry 提供了一个统一平台方便管理
4、Sentry 支持集成 Kerberos
```

缺点：

```
1、组件只支持 hive，hdfs，impala
2、不支持 hbase，yarn，kafka，storm 等
```

3.3. CDP 为什么放弃 Sentry

除了安全授权之外，Apache Ranger 还支持人性化的 Web UI，REST API 和 Auditing 等，这些都是 Sentry 所缺少的。以下我们看看 Sentry 和 Ranger 的具体区别，以了解为什么 Ranger 是 CDH 的未来选择，即 CDP。

对比	Apache Sentry	Apache Ranger	说明
认证和授权	√	√	
拒止	×	√	
Web UI	×	√	
Command Line	√	×	
HDFS Sync	√	×	
Rest API	×	√	
审计	×	√	
Impala	√	×	目前，Ranger 与 Impala 的集成还在开发中
Hive	√	√	
HDFS	√	√	Sentry 的 HDFS 支持是指同步 Hive 表的权限到 HDFS
Solr	√	√	
Kafka	√	√	
HBase	×	√	
Knox	×	√	
YARN	×	√	
Storm	×	√	
Support Tag Based	×	√	基于 Atlas 元数据 Tag 的安全策略
Row Level Filtering	×	√	
Column Masking	×	√	

如你所见，Apache Ranger 自身的功能更多，而且支持的 Hadoop 组件也更丰富，除了目前尚不支持 Impala（正在开发过程中）